LOGISTIC REGRESSION AND GENERALIZED LINEAR MODELS

W. RYAN LEE
CS109/AC209/STAT121 ADVANCED SECTION
INSTRUCTORS: P. PROTOPAPAS, K. RADER
FALL 2017, HARVARD UNIVERSITY

In this section, we introduce the idea and theory of generalized linear models, with the main focus being on the modeling aspect and capacity these models provide rather than on their inferential properties. Our approach and results are drawn from Agresti (2015) [1], which the reader is encouraged to consult for more details.

1. Linear Regression

We start with a brief overview of linear regression. Namely, we assume a dataset $\{(y_i, x_i)\}_{i=1}^n$ and consider a linear model on y_i:

$$y_i = x_i^T \beta + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently. Alternatively, in matrix form,

$$Y = X\beta + \epsilon$$

for $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Note, however, that this can equivalently be written in the form

$$Y|X, \beta \sim \mathcal{N}(X\beta, \sigma^2 I) \equiv \mathcal{N}(\mu, \sigma^2 I)$$

where we define $\mu = X\beta$. That is, we define a *linear relationship* between the mean of Y and the covariates X, determined by the parameters β. Moreover, we assume that given the covariates, all of the observations y_i are independently distributed about the linear predictor, with a symmetric Normal distribution. In particular, note that we do not necessarily need the Normality assumption; we could put a different distributional structure on ϵ and end up with a different model that is still a linear model.

2. Why Generalized Linear Models?

The above observation is key in motivating generalized linear models. In most introductions to regression, the idea of the Normal distribution being a defining feature of linear regression is deeply ingrained; however, it is not necessary to assume this, and for certain applications, it is disadvantageous to do so. For example, many real-world observations only occur on the positive real axis rather than the entirety of the reals. For such situations, one possibility would be use an Exponential/Gamma distribution on the y_i observations rather than the Normal distribution.

Such modeling considerations lead us to generalized linear models. In short, we want to keep the linear interactions between our covariates and parameters, but be able to model a more diverse range of observations than allowed by a simple linear regression model. Our observations y_i may be integer-valued, non-negative,

categorical, or otherwise unsatisfactory for a linear model, but we would still like to be able to model defining characteristics of their distribution (i.e. means, other moments, distributional parameters) using a linear relationship.

3. Natural Exponential Family

To delve further into the modeling advantages offered by generalized linear models, we consider the **natural exponential family** of distributions, of which Normal and Gamma distributions are members. Observations y are said to come from this family if they have a probability density of the form

$$f(y|\theta) = h(y)\exp(y^T\theta - b(\theta))$$

where θ is called the **natural parameter** of the distribution.

Now consider the log-likelihood, which is the fundamental quantity for statistical inference.

$$l(\theta) = \log f(y|\theta)$$

Two important identities regarding the log-likelihood concern the expectations of its derivatives, known as the **score function**

$$S(\theta) \equiv \frac{\partial l}{\partial \theta}$$

Note, in particular, that the maximum likelihood estimator is simply the root of the *score equation*

$$S(\hat{\theta}) = 0$$

Proposition 3.1. *Let $l(\theta)$ be the log-likelihood and $S(\theta)$ be the score function. Then, under suitable regularity conditions allowing for differentiation under the integral, the following identities hold.*

$$E[S(\theta)] = E\left(\frac{\partial l}{\partial \theta}\right) = 0$$

$$I(\theta) \equiv -E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = E\left(\frac{\partial l}{\partial \theta}\right)^2 = var(S(\theta))$$

*That is, the **Fisher information matrix** is the variance of the score function.*

Proof. Letting $f(y|\theta)$ denote the density,

$$S(\theta) = \frac{\partial}{\partial \theta}\log f(y|\theta) = \frac{\partial_\theta f(y|\theta)}{f(y|\theta)}$$

where ∂_θ denotes the partial derivative with respect to θ, and so the expectation is

$$E[S(\theta)] = \int_y \frac{\partial_\theta f(y|\theta)}{f(y|\theta)} f(y|\theta)dy = \int_y \partial_\theta f(y|\theta)dy = \partial_\theta \int_y f(y|\theta)dy = 0$$

where we use the regularity condition allowing us to take the derivative out of the integral, and note that the integral is always equal to unity by the fact that the integrand is a probability density.

For the second identity, note that the first equality is the definition of the Fisher information, and the last identity is true because the expectation of the score is

zero, as we just showed. Thus, the only identity to be proved is the middle identity. We can express the left-hand side as

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = \int_y \frac{\partial^2 l}{\partial \theta^2} f(y|\theta) dy = \partial_\theta l(\theta) \partial_\theta f(y|\theta)|_{\theta=-\infty}^{\infty} - \int_y \partial_\theta l(\theta) \partial_\theta f(y|\theta) dy$$

where the second equality follows from integration by parts. The first term is zero for suitable regularity conditions. We now use the expression for the score function to write

$$\partial_\theta f(y|\theta) = \partial_\theta l(\theta) \cdot f(y|\theta)$$

and so the expression becomes

$$-E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = \int_y (\partial_\theta l(\theta))^2 f(y|\theta) dy = E\left(\frac{\partial l}{\partial \theta}\right)^2$$

as desired. □

Using the above proposition, we find the following useful relations regarding the terms in the exponential family density.

$$\mu \equiv E[y|\theta] = b'(\theta)$$

$$var(y|\theta) = b''(\theta)$$

In other words, $b(\theta)$ is the **cumulant function** of the distribution.

Note in particular that the Poisson, Bernoulli, and Normal distributions are members of the natural exponential family. For example, we can write the probability mass function of a single Poisson observation as

$$f(y|\mu) = \frac{\mu^y e^{-\mu}}{y!} = (y!)^{-1} \exp(y \log \mu - \mu)$$

which has the form of the exponential family with $\theta = \log(\mu)$ and $b(\theta) = \exp(\theta)$. Similar derivations can be made for the Bernoulli and Normal cases, among others.

An important point to note about exponential family distributions is that they are completely characterized by the relation between the mean and the variance.

Theorem 3.2. *Let $f(y|\theta)$ be a density in the natural exponential family. Then, the variance of the distribution can be written in terms of its mean μ; that is,*

$$var(y|\theta) = v(\mu)$$

for some function v. Moreover, the function v uniquely specifies the distribution f.

For example, the Poisson distributed can be *defined* as the exponential family distribution such that

$$\text{var}(y|\mu) \equiv v(\mu) = \mu$$

and similarly for the Bernoulli and Normal distributions. In other words, there is only one exponential family distribution satisfying the relation $v(\mu) = \mu$, and this is the Poisson distribution.

4. Logistic Regression

Let us now consider the case of **logistic regression** to see an example in which the linear model is entirely inadequate. We now assume *binary data*, in which y take only the values $0, 1$. The model is typically given by

$$\text{logit}(\mu) = x^T\beta$$

which implicitly models the observations y as

$$y|\mu \sim \text{Bern}(\mu)$$

independently. That is, the observations given the probability μ (or mean) are independently drawn from a Bernoulli distribution, each with its own probability of success μ. These μ are related to the covariates x through a linear predictor of their logit:

$$\text{logit}(\mu) \equiv \log(\mu/(1-\mu))$$

Putting this all together, we can write the model as:

$$y|x, \beta \sim \text{Bern}\left(\frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}\right) = \text{Bern}(\text{logit}^{-1}(x^T\beta))$$

where $\text{logit}^{-1}(x) = e^x/(1+e^x)$ is the inverse of the logit function defined above.

The logit can be shown to be a reasonable function *linking* the linear predictor to the mean in a number of ways. When the covariates are also given a distribution and assumed to come from a Normal distribution, then the posterior distribution of y is precisely given by the logistic regression equation above. In addition, when the Bernoulli distribution is written in the exponential family form, one can see that the natural parameter is precisely the logit, as we show below.

One interesting property of logistic regression models for classification is implied by their roots in *linear* models. Often, for classification purposes, one would like to predict the class (0 or 1) of a new sample $\tilde{y}$ based on a model fit on the existing data. In this case, we would like to predict $\tilde{y} = 1$ if

$$P(\tilde{y} = 1|\tilde{x}, \beta) \equiv \tilde{\mu} \equiv \frac{\exp(\tilde{x}^T\beta)}{1+\exp(\tilde{x}^T\beta)} \geq c$$

for some constant c, generally $c = 1/2$.

Proposition 4.1. *The predictive classification of $\tilde{x}$ based on the criterion*

$$P(\tilde{y} = 1|\tilde{x}, \beta) \geq c$$

*for any $c \in (0, 1)$ is equivalent to the **linear discriminant** or **linear decision boundary** given by*

$$\tilde{x}^T\beta \geq \textit{logit}(c)$$

Proof. We first prove that the inverse logit is a monotonically increasing function.

$$\partial_x \text{logit}^{-1}(x) = \frac{e^x(1+e^x) - e^x \cdot e^x}{(1+e^x)^2} = \frac{e^x}{(1+e^x)^2} > 0$$

for any finite value of x. Thus, since the inverse logit function is a monotonically increasing function of its argument, it is one-to-one, an inverse exists (the logit), and the inequality $\text{logit}^{-1}(x) \geq c$ is equivalent to $x \geq \text{logit}(c)$. □

Note that as its name suggests, the decision boundary is linear; it is simply a hyperplane in p-dimensional space, where p is the number of predictors. As shown by our derivation, this entirely results from the linear model we assumed in forming the logistic regression model. In other words, because we used a linear predictor $\mu = x^T\beta$, the decision boundary is also linear. Had we used a nonlinear predictor, the boundary would have correspondingly changed as well.

5. Poisson Regression

Another case in which the linear model fails to provide an adequate specification is when the data are assumed to be *count-based*, so that they only take values in the natural numbers. In this case, the model is typically given by

$$\log(\mu) = x^T\beta$$

which again implicitly models the observations as

$$y|\mu \sim \text{Pois}(\mu)$$

or, more compactly, as

$$y|x, \beta \sim \text{Pois}(\exp(x^T\beta))$$

Note that there is no intrinsic need to model the log of the mean linearly, aside from the fact that the linear predictor may be negative. Thus, if we are confident that the linear predictor will take only positive values, we can also simply consider

$$y|x, \beta \sim \text{Pois}(x^T\beta)$$

This simply adjusts the interpretation of the parameters β from a multiplicative effects to additive ones, as in the linear regression model.

Moreover, it is often the case that the mean of the observation y may be proportional to an *offset* t. For example, if y is the number of malfunctions for a gadget, it may very well be proportional to the time t that the gadget has been in operation. In this case, we can model the mean as

$$\log(\mu/t) = x^T\beta \Rightarrow \log(\mu) = \log(t) + x^T\beta$$

In other words, this is equivalent to using the log of the offset as a covariate and forcing the coefficient to be 1.

It is important to note that a defining limitation of using a Poisson model for count data is that the distribution inherently limits the variance to be equal to the mean:

$$\text{var}(y|\mu) = \mu = E[y|\mu]$$

For *over-dispersed* data, this can be a debilitating limitation. For such cases, there are many modeling alternatives that can still provide principled solutions. One popular approach is to use a **negative binomial** distribution, which provides an additional degree of freedom to model the variances accordingly. For data that contains many zeros (more than a Poisson distribution would predict), another alternative is to use a **zero-inflated Poisson** (ZIP) model, which puts a desired amount of mass on the 0 observation to better model the data.

6. General GLMs

As mentioned in the introduction, the linear, logistic, and Poisson regression models are special cases of a more general framework. Now proceeding to full generality, a **generalized linear model** (GLM) consists of three components, namely

- **Random component**. This defines the distribution of y given its parameters (its mean, other moments). This distribution is generally within the natural exponential family.
- **Linear predictor**. Denoted by η, the linear predictor defines the relationship between the covariates x, parameters β, and the mean of the distribution of y.
$$\eta \equiv x^T\beta$$
- **Link function**. The link function g defines the relationship between the linear predictor η and the mean μ.
$$g(\mu) = \eta = x^T\beta$$

The key point to understand about GLMs is the separation between the random component and the link function. In specifying a GLM, the data scientist has full rein over what random component for y is most plausible for your research question. Moreover, one can also specify the link function that makes the linear relationship between $g(\mu)$ and x most reasonable and plausible, based on domain knowledge. This is in contrast to many other types of models, in which (for various purposes, such as ease of optimization) the choice of one of these aspects determines the others.

We now show that the various regression models we have considered are, in fact, special cases of the above framework.

Example 6.1. For the linear model, we simply use
$$y|\mu \sim \mathcal{N}(\mu, \sigma^2)$$
as the random component, assuming σ^2 is known, and $g(\mu) = \mu = \eta$ as the link function; that is, the identity is the link function for the linear model. □

Example 6.2. For Poisson regression, we use
$$y|\mu \sim \text{Pois}(\mu)$$
as the random component and $g = \log$ as the link function. □

Example 6.3. Finally, for the logistic regression, we use
$$y|\mu \sim \text{Bern}(\mu)$$
and use $g(\mu) = \log(\mu/(1-\mu))$ as the link function (i.e. the logit). □

For each of these cases, when the random component lies in the natural exponential family, there exist link functions that are, in a sense, most natural for the problem at hand. These link functions are known as the **canonical link functions**. Namely, the canonical link sets the natural parameter equal to the linear predictor,
$$\theta \equiv x^T\beta$$
Thus, investigating the relationship between θ and μ yields the link function, as shown in the following proposition.

Proposition 6.4. *The canonical link functions are precisely the log link for the Poisson, logit link for the Bernoulli, and identity link for the Normal linear model.*

Proof. We prove the result for the logit link, as it is the most involved, and leave the remaining link functions to the reader. Note that the Bernoulli mass function can be written as

$$f(y|\mu) = \mu^y(1-\mu)^{1-y} = \exp[y\log(\mu) + (1-y)\log(1-\mu)]$$
$$= \exp\left[y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right]$$

Thus, we see that the natural parameter is

$$\theta = \text{logit}(\mu) = x^T\beta$$

where the last equality is required for the link to be a canonical link function. This proves that the logit link is the canonical link function for the Bernoulli. □

The beauty of encompassing all of these models under the GLM framework is that there is both theory and inferential methods associated with general GLMs that can easily be applied in each specific case. For example, there exist general asymptotic results for parameters estimated based on GLM models, which we can apply to the specific model under consideration.

Theorem 6.5. *If $\hat{\beta}$ is the maximum likelihood estimator for a GLM with linear predictor $\eta = X\beta$, then:*

$$\hat{\beta} \to_{\mathcal{L}} \mathcal{N}(\beta, (X^TWX)^{-1})$$

where $\to_{\mathcal{L}}$ represents convergence in distribution, and where W is an appropriate weight matrix.

Thus, in general, the maximum likelihood estimators have an asymptotic Normal distribution, which we can use to construct asymptotic confidence intervals when doing inference. Moreover, we have the generic **likelihood equations** for a GLM, which provides parameter inference equations for any GLM.

Theorem 6.6. *If $l(\beta)$ is the log-likelihood under a GLM with linear predictor $\eta = X\beta$, then the following likelihood equations hold,*

$$\frac{\partial l(\hat{\beta})}{\partial \beta} = X^TDV^{-1}(y-\mu) = 0$$

where D is the diagonal matrix of terms $\frac{\partial \mu_i}{\partial \eta_i}$ and V is the diagonal matrix of variances of each observation.

Note that in the above result, β is implicitly controlled by this equation through μ. Thus, there is no need to consider first-order conditions and optimality results every time we consider a new GLM. Once we have shown that the model under consideration is, in fact, a member of the GLM family, we can simply compute the corresponding matrices D, V and solve the likelihood equations to compute the maximum likelihood estimator for β.

References

[1] A. Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley, 2015.